Beilstein-Institut

Glyco-Bioinformatics – *Bits 'n' Bytes of Sugars*
October 4th – 8th, 2009, Potsdam, Germany

# Automated N-Glycan Composition Analysis with LC-MS/MSMS

**Hannu Peltoniemi**[1,*], **Ilja Ritamo**[2], **Jarkko Räbinä**[2] **and Leena Valmu**[2]

[1]Applied Numerics Ltd, Nuottapolku 10 A8, FI-00330 Helsinki, Finland.

[2]Finnish Red Cross Blood Service, R&D, Kivihaantie 7, FI-00310 Helsinki, Finland.

**E-Mail:** *hannu.peltoniemi@appliednumerics.fi

*Received: 1st March 2010/ / Published: 10th December 2010*

## Abstract

Compared to proteomics the mass spectrometric glycan analysis still employs a lot of manual work and the differential glycomics can be a burden with increasing number of spectra. Our aim is to ease these tasks by using in-house developed glycomic software in combination with existing proteomics tools. The resulting workflow is targeted especially to glycan LC-MS/MSMS analytics and can be run with a minimal amount of human intervention. Here the method was applied to cell surface N-glycans from umbilical cord blood derived mononuclear cells. The final goal is to profile and differentiate the stem cell surface glycans which are being analysed at the Finnish Red Cross Blood Service.

## Background

Traditionally mass spectrometric (MS) glycan analysis [1, 2] has mainly been performed by one-dimensional matrix assisted laser desorption ionization (MALDI) –– time-of-flight (TOF) analysis, whereas in proteomics the use of liquid-chromatography (LC) coupled to electrospray (ESI)-MS has increased rapidly in recent years. For glycans there are many good reasons to favour MALDI, including simpler one-dimensional spectra, established wet lab procedures, existing software etc. The benefit of the LC is on the other hand in additional glycan separation, which permits for example the isomeric differentiation of glycans [3, 4].

Tandem mass spectrometric (MSMS) fragmentation analysis is also more feasible to perform on ESI-MS instrument, although it is performable today also on some MALDI instruments. The use of LC-MS/MSMS has been limited both by the complexity of the spectra, namely in the form of multiple different charge states and metal adducts, and by expensive instrumentation. Also, a major drawback is the lack of suitable software to ease the LC-MS/MSMS data analysis. Even though the available glycan software is limited, that is not the case with proteomics. A lot of software to analyse peptide LC-MS/MSMS data exists also as open source.

In the R&D department of the Finnish Red Cross Blood Service the focus is on the cell surface glycoconjugates from human cells aimed for cellular therapy. The MALDI-TOF glycan profiling of different stem cell classes, including embryonic [5], hematopoietic [6] and mesenchymal [7] cells, has previously been performed. Lately, more focused cell surface glycoconjugate analytics has been developed. Here, the cell surface proteins are biotinylated [8] and glycans are released from them. Within the workflow the glycans are further reduced in order to eliminate anomer peaks and permethylated in order to increase the ionizability in ESI and to ease the interpretation of fragmentation data [9]. The glycan samples are analyzed by reverse phase (RP)-nano-LC (LC Packings Dionex) coupled to LTQ Orbitrap XL (ThermoFisher Scientific) MS with an ESI ion source.

The experimental raw data can be represented as a LC-MS 2D map which has MSMS spectra embedded (Fig. 1). The aim of glycan identification is to find the set of glycans that *explains* the data *best*, and to calculate the total intensity (profile) for the identified glycans. Prior to the glycan identification additional pre-processing steps, for example peak picking, deisotoping and feature detection, are required.

The existing glycomic software was reviewed, but no single software to solve the overall problem was found. Among the most prominent glycan software were GlycoWorkbench [10] and Glyco-Peakfinder [11]. However, they have been developed for the analysis of one spectrum at a time and cannot be automated for a larger set of spectra acquired by LC-MS/MSMS analysis. Also, at the starting time of this project (summer 2008) the software was not yet published as open source. Other available glycan mass spectrometry related software includes free GlycoMod web tool [12] and proprietary SimGlycan [13].

For peptide LC-MS analysis there exists plenty of software both as open source (for example msInspect [14] and OpenMS [15]) and as proprietary ones (for example Progenesis LC-MS by Nonlinear Dynamics Ltd [16] and DeCyder by GE Healthcare Ltd [17]). Peptide LC-MS analysis differs slightly from the corresponding glycan analysis, mainly in the form of several charge carriers, not just hydrogen. But there are also many similar attributes in the analysis of these two analytes, peptides and glycans, including feature detection, alignment and feature comparison between different samples.

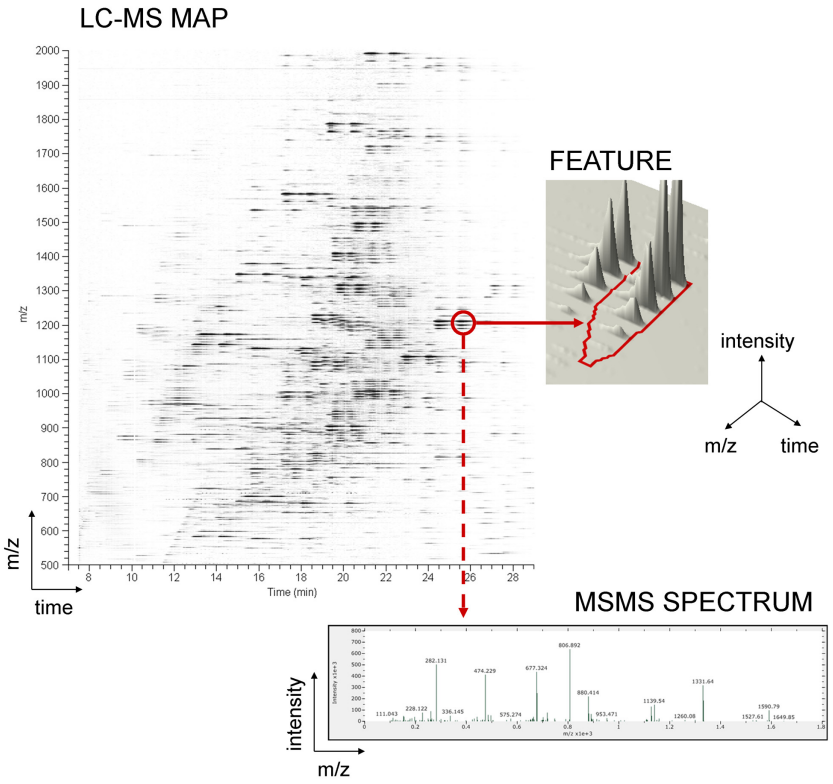


**Figure 1.** The LC-MS/MSMS experiment data represented as LC-MS map of eluted MS features and MSMS spectra generated by CID fragmentation of glycans.

After the survey on glycoinformatic tools the conclusion was to tailor our own glycomic software, but to apply as much of the existing software as possible, especially the tools developed for proteomics, as part of the workflow. The emphasis was more on automation and less on interactive use and user interfaces. The identification of glycan compositions was set to be sufficient to start with. To enable rapid prototyping the in-house developed part of the software was decided to be done mostly with the R statistical computing environment [18]. The R is open source software containing a lot of numerical and statistical methods, including bioinformatics methods [19].

## The Glycan Identification Workflow

The glycan LC-MS/MSMS identification workflow (Fig. 2) combines existing proteomic software and in-house developed glycan specific tools. The analysis starts from a 2D LC-MS map containing chromatographic (retention time) and mass (m/z) dimensions together with embedded tandem MSMS spectra created by fragmentation of glycans. The result is a profile of matching glycans and a suggestion of the simplest set of glycans that can explain the measured data. The glycan specific tools (steps 3 – 6 in the workflow) are based on the in-house developed R library called *Glycan ID*. The library methods include spectrum matching, outlier removal, statistical scoring and visualization. The aim of the library is to enable fast development of new workflow variants when new requirements appear.

The tools in the workflow are:

**1. Identify Features**

Potential glycan features are identified with Progenesis LC-MS (Nonlinear Dynamics Ltd) [16] software developed originally for peptide analysis.

**2. Extract MSMS Spectra**

MSMS spectra with identified charge states (deisotoped) is extracted with Mascot Distiller (Matrix Science Ltd) [20]. The software has originally been developed as a pre-processor for protein identification search engine Mascot (Matrix Science Ltd).

**3. Match Compositions (MS)**

The glycan compositions which match to feature masses are searched. The feature matching is done either against theoretical compositions generated *de novo* with a given set of rules [21], or against a user given list of glycan compositions (database). Several charge carrier ion types and neutral adducts can be used. Outliers can be removed by iteratively applying linear fitting and elimination of compositions with a mass difference greater than two standard deviations. The tool uses an approach which is very similar to the one used by Glyco-Peakfinder [11] or GlycoMod [12].

**4. Match Compositions (MSMS)**

Glycan compositions which match the precursor masses and MSMS fragment spectra are searched. The precursor compositions are found as above. Fragment matching is done either against all theoretical fragments that any glycan structure with a given composition could produce [21, 22], or against theoretical or measured spectra in a given MSMS spectrum database. Outlier matches can be removed as above. The matched compositions are ranked by a statistical score defined by a logarithm of a product of two probabilities:

1) The probability that a random set of fragments would have as many or more shared peaks with the measured spectrum as the ranked composition [21] and, 2) The probability that by randomly selecting the observed number of shared peaks the same or higher amount of intensity can be covered. Two optional filtering steps are included: 1) An MSMS spectrum is taken into account only if any mass difference between two peaks matches a list of given masses, typically composed by one or two monosaccharide masses. 2) To ease the differentiation between N-acetyl-neuraminic acid (Neu5Ac) and N-glycolyl-neuraminic acid (Neu5Gc), a given residue is allowed to exist in a proposed composition only if the MSMS spectrum contains at least one of the given marker ions derived from these sialic acids.

**5. Combine MS and MSMS**

The results of MS and MSMS matching are combined so that the MSMS identification is included in the specific MS feature if the mass and retention time differences between the MS feature and MSMS precursor are less than the given tolerances.

**6. Deconvolute**

The last fully automated step in the workflow is the calculation of the total intensity and score for each proposed glycan by summing the measured feature intensities and MSMS scores with different charge states and charge carrier types, namely metal adducts and protons. The glycans are further grouped so that the proposed compositions matching a common set of features are categorized into the same group. As these sets are independent, the analysis of one group does not have an effect on the analyses of other groups. For each group, one glycan composition is marked as most likely the correct one if there is only one composition that matches all group features and if the composition has the highest score. Otherwise the group is marked to be contradictory. The glycan profile is created from the deconvoluted data and the possible contradictory groups are manually resolved based on the biological information available.

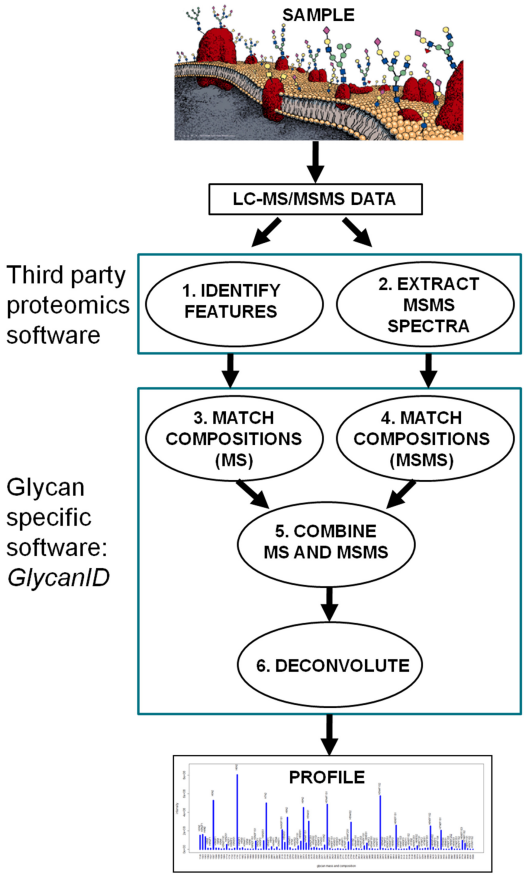


**Figure 2.** The glycan identification workflow.

## Example: N-glycans from Umbilical Cord Blood Mononuclear Cells

The workflow was applied to the cell surface N-glycans from umbilical cord blood derived mononuclear cells. The total cellular N-glycome from the same cell type has previously been analysed by MALDI-TOF [6].

The cell surface proteins were labelled with biotin and enriched by streptavidin coupled magnetic beads as previously described [23]. N-glycans were released from the cell surface protein fraction by PNGase F and reduced with NaBH4. The reduced N-glycans were permethylated as described in [24]. The permethylated and reduced N-glycans were loaded into RP precolumn (Atlantis dC 18, Waters) and separated in analytical RP column (PepMap 100, Dionex Corporation). Ultimate 3000 LC instrument (Dionex Corporation) was operated

in nano scale with a flow rate of 0.3 µl/min. The eluted glycans were introduced to LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific Inc.) via ESI Chip interface (Advion BioSciences Inc.) in the positive-ion mode.

The glycan profiling was performed against *de novo* generated glycan compositions allowing the following restrictions: 1) Monosaccharides with 3 – 15 hexoses (H), 2 – 15 N-acetylhexosamines (N), 0 – 6 deoxy-hexoses (F) and 0 – 6 N-acetyl-neuraminic acids (S), 2) Charge carriers as either sodium or hydrogen adducts and 3) Assumption of the intact N-glycan core. Mass tolerance was set to 5 ppm with MS and to 10 ppm with MSMS spectra. The workflow started with approximately 1000 MS features and 700 MSMS precursors and ended with 54 different glycan compositions proposed by automated identification and further classified manually as biologically credible ones. About 40% of all the features had at least one matching composition, whereas the number was 90% for the 1/10 of the highest intensity features with a charge state two or higher. Naturally, the match coverage would be higher if the number of accepted monosaccharide residues and metal adducts had been larger, but also the probability to get false interpretations would increase. The future challenge will be to tune the analysis so that both the sensitivity and selectivity are optimized.

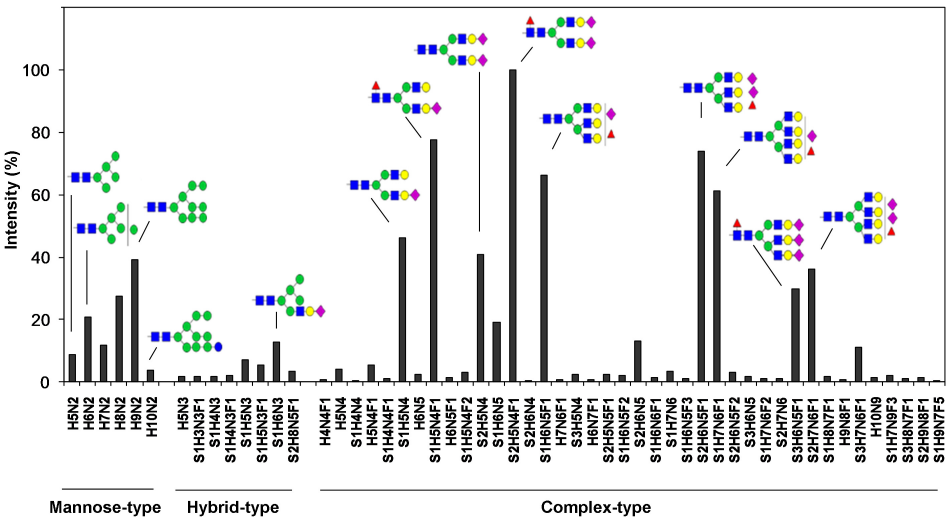


**Figure 3.** Cell surface N-glycan composition profile of umbilical cord blood derived mononuclear cells. The added glycan structures are based on educated guess of possible structures matching the identified compositions.

The calculated N-glycan compositions (Fig. 3) fit very well into our previously published glycan structure data of cord blood derived mononuclear cells [6], but clear differences in the cell surface N-glycan profile are seen in comparison with the total cellular N-glycan

profile. On the cell surface far less high mannose-type glycans are observed, whereas complex-type glycans seem to predominate. Also fucosylated and sialylated structures are heavily enriched on the cell surface N-glycans.

## Conclusions

In the mass spectrometric glycan analysis the number of features using two-dimensional LC-MS methodology is far greater than in other analytical methods typically performed in one dimension. In LC-MS analysis the complexity is increased both by ESI, which produces multiple charged ions, as well as by second dimension, introduced by chromatographic retention time. Additional complexity is still involved in the number of different metal adducts detected in the glycan analyses. Therefore, in order to utilize the additional possibilities that LC-MS analysis introduces to the glycan structure determination, a competent data handling tool is essential in order to simplify the otherwise extremely laborious interpretation of the data. However, by the limited number of monosaccharide residues allowed in the calculation and by the overlapping composition masses, some level of non-uniqueness will always be present within the given result. If an unambiguous glycan composition is required, a manual verification by an expert in glycobiology is definitely needed. The advantage of the automated glycan identification software is that it can easily show the possibilities and can generate a suggestion of a simple solution.

The glyco-bioinformatics is an emerging branch of informatics with some developed software but still with many application areas to be covered. The development of novel and automated applications could speed up if the existing software could be used as part of a novel workflow. To enable this, the software should be developed so that it can be run without a user interface, as a batch process, a library or a web service. Naturally, open source software would be the most beneficial, but proprietary ones are not excluded assuming there are no other obstacles with the workflow use. When the software presented in this study is matured enough for publication it is planned to be opened for wider use either as a web service or as open source software.

## Abbreviations

| | |
|---|---|
| ESI | Electrospray ionisation |
| F | deoxyhexose (fucose) |
| H | Hexose |
| LC | Liquid chromatography |
| MALDI | Matrix-assisted laser desorption ionisation |
| MS | Mass Spectrometry |
| MSMS | Tandem Mass Spectrometry |
| N | N-acetyl hexosamine |
| Neu5Ac | N-acetyl-neuraminic acid |
| Neu5Gc | N-glycolyl-neuraminic acid |
| RP | Reversed phase |
| S | Neu5Ac (sialic acid) |
| TOF | Time-of-flight |

## References

[1] North, S.J., Hitchen, P.G., Haslam, S.M., Dell, A. (2009) Mass spectrometry in the analysis of N-linked and O-linked glycans. *Curr. Opin. Struct. Biol.* **19**:498 – 506. doi: http://dx.doi.org/10.1016/j.sbi.2009.05.005.

[2] Zaia, J. (2008) Mass spectrometry and the Emerging Field of Glycomics. *Chem. Biol.* **15**:881 – 892. doi: http://dx.doi.org/10.1016/j.chembiol.2008.07.016.

[3] Wuhrer, M., Deelder, A.M., Hokke, C.H. (2005) Protein glycosylation analysis by liquid chromatography – mass spectrometry. *J. Chrom. B* **825**:124 – 133. doi: http://dx.doi.org/10.1016/j.jchromb.2005.01.030.

[4] Ruhaak, L.R., Deelder, A.M., Wuhrer, M. (2009) Oligosaccharide analysis by graphitized carbon liquid chromatography – mass spectrometry. *Anal. Bioanal. Chem.* **394**:163 – 174. doi: http://dx.doi.org/10.1007/s00216-009-2664-5.

[5] Satomaa, T., Heiskanen, A., Mikkola, M., Olsson, C., Blomqvist, M., Tiittanen, M., Jaatinen, T., Aitio, O., Olonen, A., Helin, J., Hiltunen, J., Natunen, J., Tuuri, T., Otonkoski, T., Saarinen, J., Laine, J. (2009) The N-glycome of human embryonic stem cells. *BMC Cell Biol.* **10**:42.
doi: http://dx.doi.org/10.1186/1471-2121-10-42.

[6] Hemmoranta, H., Satomaa, T., Blomqvist, M., Heiskanen, A., Aitio, O., Saarinen, J., Natunen, J., Partanen, J., Laine, J., Jaatinen, T. (2007) N-glycan structures and associated gene expression reflect the characteristic N-glycosylation pattern of human hematopoietic stem and progenitor cells. *Exp. Hematol*. **35**:1279 – 1292.
doi: http://dx.doi.org/10.1016/j.exphem.2007.05.006.

[7] Heiskanen, A., Hirvonen, T., Salo, H., Impola, U., Olonen, A., Laitinen, A., Tiitinen, S., Natunen, S., Aitio, O., Miller-Podraza, H., Wuhrer, M., Deelder, A.M., Natunen, J., Laine, J., Lehenkari, P., Saarinen, J., Satomaa, T., Valmu, L. (2009) Glycomics of bone marrow-derived mesenchymal stem cells can be used to evaluate their cellular differentiation stage. *Glycoconj. J.* **26**:367 – 384.
doi: http://dx.doi.org/10.1007/s10719-008-9217-6.

[8] Elia, G. (2008) Biotinylation reagents for the study of cell surface proteins. *Proteomics* **8**:4012 – 4024.
doi: http://dx.doi.org/10.1002/pmic.200800097.

[9] Costello, C., Contado-Millera, J.M., Cipollo, J.F. (2007) A glycomics platform for the analysis of permethylated oligosaccharide alditols. *J. Am. Soc. Mass Spectrom.* **18**:1799 – 1812.
doi: http://dx.doi.org/10.1002/pmic.200800097.

[10] Ceroni, A., Maass, K., Geyer, H., Geyer, R., Dell, A., Haslam, S.M. (2008) GlycoWorkbench: A Tool for the Computer-Assisted Annotation of Mass Spectra of Glycans. *J. Proteome Res.* **7**:1650 – 1659.
doi: http://dx.doi.org/10.1002/pmic.200800097.

[11] Maass, K., Ranzinger, R., Geyer, H., von der Lieth, C-W., Geyer, R. (2007) "GlycoPeakfinder" – *de novo* composition analysis of glycoconjugates. *Proteomics* **7**:4435 – 4444.
doi: http://dx.doi.org/10.1002/pmic.200700253.

[12] Cooper, C.A., Gasteiger, E., Packer, N.H. (2001) GlycoMod – A software Tool for Determining Glycosylation Compositions from Mass Spectrometric Data. *Proteomics* **1**:340 – 349.
doi: http://dx.doi.org/10.1002/1615-9861(200102)1:2<340::AID-PROT340>3.3.CO;2-2.

[13] SimGlycan http://www.premierbiosoft.com/glycan/index.html

[14] May, D., Law, W., Fitzgibbon, M., Fang, Q., McIntosh, M. (2009) Software Platform for Rapidly Creating Computational Tools for Mass Spectrometry-Based Proteomics. *J. Proteome Res.* **8**:3212 – 3217.
doi: http://dx.doi.org/10.1021/pr900169w.

[15] Kohlbacher, O., Reinert, K. (2009) OpenMS and TOPP: Open Source Software for LC-MS Data Analysis. In: *Proteome Bioinformatics*, ed. by Simon J. Hubbard and Andrew R. Jones, Methods in Molecular Biology. Humana Press., vol. 604, chap. 14.

[16] Progenesis LC-MS, http://www.nonlinear.com

[17] DeCyder MS Differential Analysis Software, http://www.gelifesciences.com

[18] The R Project for Statistical Computing, http://www.r-project.org.

[19] BioConductor, http://www.bioconductor.org.

[20] Mascot Distiller, http://www.matrixscience.com

[21] Joenväärä, S., Ritamo, I., Peltoniemi, H., Renkonen, R. (2008) N-Glycoproteomics – an automated workflow approach. *Glycobiology* **18**:339 – 349.

[22] Peltoniemi, H., Joenväärä, S., Renkonen, R. (2009) De novo glycan structure search with the CID MS/MS spectra of native N-glycopeptides. *Glycobiology* **19**: 707 – 714.

[23] Scheurer, S.B., Rybak, J-N., Roesli, C., Brunisholz, R.A., Potthast, F., Schlapbach, R., Neri, D., Elia, G. (2005) Identification and relative quantification of membrane proteins by surface biotinylation and two-dimensional peptide mapping. *Proteomics* **5**:2718 – 2728.

[24] Kang, P., Mechref, Y., Klouckova, I., Novotny, M.V. (2005) Solid-phase permethylation of glycans for mass spectro-metric analysis. *Rapid Commun. Mass Spectrom.* **19**:3421 – 3428.